4-22-2021

Social tags and Controlled Vocabulary: A Comparative Study of Documents in Chemistry and Literature

Dr Indira B
indirab575@gmail.com

Social tags and controlled vocabulary: A comparative study of documents in chemistry and literature

Dr Indira B
Indirab575@gmail.com

Abstract

The online information retrieval process enables users to assign keywords for easy access of documents. These keywords popularly known as social tags have now become a popular trend in the information organization. Many online library catalogues has provided provision which allows users to add tags to documents. These social tags has many similarities and variations with the controlled vocabularies used in libraries. Moreover the usage of social tags varies in each subject. The present study tries to compare the social tags and controlled vocabularies in literature and chemistry. The analysis have proved that social tags that exactly match with the subject headings are very high in chemistry and very poor in literature . The study has also proved that the total number of social tags are high in literature and less in chemistry and the social tags are more suitable in fiction than nonfiction.

Keywords :- social tags, controlled vocabulary, literature, chemistry, fiction, nonfiction.

Online library catalogues provide keywords for easy and fast access of documents. These key words are usually subject descriptions of the content which they derive using standard thesaurus and vocabularies like Library of congress subject headings, Sears' list of subject heading, MeSH etc. These subject headings act as keywords to the content of document through which user access to resources. A restrictive and selective number of experts were involved in the

conception, assignment and maintenance of these vocabularies. Information scientists and professionals in library science are involved in assigning subject headings to documents after perusing its contents. The work which involved so much of expertise is now being done by users. When online sources became widely used and OPAC became the access points to resources in libraries, a detailed subject access to documents became a vital need.

The interactive web which inculcates features such as open sharing and collaboration led to the advent of social media and networking sites where the users can describe their own or other's resources mostly for personal purposes. These resources were mainly photographs, songs and videos and the descriptions given by users came to be known as tags. Then social bookmarking cites such as Delicious, CiteULike (www.citeulike.org), photo-sharing systems, such as Flickr(www.flickr.com) Instagram (www.instagram.com), Twitter (https://twitter.com) and OPAC systems, such as BiblioCommons (www.bibliocommons.com) originated and gained popularity in the web world. Many social cataloguing cites such as Goodreads (https://www.goodreads.com), Litsy (https://www.litsy.com), Anobii (https://www.anobii.com), Readgeek (https://www.readgeek.com) and LibraryThing (https://www.librarything.com) were also active on the web.

The Library Thing which was developed by Tim Spalding in 2005 is a website for social cataloguing which allows any individual or organization to catalogue any number of books at no cost. In order to open an account only a user name which will be the email of the user and a password is required. The members can also edit their profile to make a private account so that others cannot view what books are catalogued. Library Thing has access to the Library of congress, six national Amazon sites and more than two thousand libraries around the world. It allows classifying books using DDC and LCC, tag books and even scan books using Library Thing iOS App. Now Library Thing has 25, 91,219 members and 15, 40, 84,495 books catalogued and this number is increasing day after day. It also offers social networking space where members can interact with other similar libraries and put reviews about books. It has an icon called Zeitgeist which provides vital statistics about members, books catalogued, most reviewed books, top authors, largest libraries, top tags and top series. As library Thing has a bulky list of catalogued books it is assumed to be significant for this study.

Cataloguing of book is very simple in Library Thing, Click on "Add books" icon on home page, give the available details as to author and title, the cataloguing details of similar title books will appear on right side of the screen and can be easily selected. Once added, the publication details, image of the book, members who have already added this book, social tags, recommendations and

reviews are displayed. There are no restrictions to the number of tags included into library Thing. The clusters of tags are enumerated as tag clouds. The tag clouds are the common way through which tags present in the system is visualized. Here the tags listed alphabetically with larger, bolder and prominent fonts are more popular tags. It helps to identify other titles that have assigned similar tags.

OCLC classification provides access to most commonly used classification numbers in DDC/LCC and subject headings for individual works. It provides searches under author, title, ISBN, ISSN and Subject keywords. The fast subject headings are controlled vocabulary invented by OCLC as a simplification of Library of Congress Subject heading[2].Many libraries depend on OCLC Classify for call numbers and subject Headings for easy and fast technical organization of their collection, hence it is selected for a comparative analysis.

Objectives

1. To identify the percentage of social tags similar (exact match) to subject headings given in OCLC Classify
2. To analyze the percentage of tags other than subject terms (user related tags) used for identifying the documents
3. To compare the total number of social tags and subject headings in literature and chemistry
4. To find out the social tag that has high frequency of use.

5. To illustrate the application of social tags in fiction and non fiction

Related studies

Vaidya and Harinarayana (2016) have discovered twenty high frequency words by using cosine similarity measure. They took 100 titles from library and information science and collected social tags from library thing and controlled vocabularies from library of congress subject headings. They first noticed that large number of tags are assigned to a book when compared to subject headings hence duplicate entries were removed and the final list of terms were arranged in descending order based on frequency. Cosine similarity measure was applied to find the similarity coefficient the result of this mathematical analysis proved that vocabulary applied by users in the form of tags is less similar to controlled vocabulary. Their conclusion states that social tags can never replace the value of controlled vocabularies in the context of information retrieval. They have also suggested that library discovery systems should include provisions for adding social tags and later these tags should be incorporated into the information retrieval system. The same authors have conducted a similar research on social tags in Marine science domain, (2018). Here they have tried to find out the relevancy of social tags in retrieval of information through TF-IDF statistical tool and Jaccard Similarity Test. The common words present in full text and social tags are considered for study. The study tries to find solution to the questions as to what

extend the users assign words similar to authors as tags and whether weight can be assigned to terms and provide relevant ranking without knowing the semantic meaning. The authors have proved using Jaccard similarity calculation that social tags are less in similarity to the words assigned by authors and these social tags though found useful may not replace the other varieties of structured vocabularies and they applied TF-IDF to prove that it possible to find significant words in terms of relevance or weightage.

Rahman (2012) suggested a hybrid catalogue of social tags and controlled vocabulary for best information retrieval. He devised a coding system to pull out the similarities and dissimilarities in the social tags of 20 science books in Library Thing with Library of Congress subject headings. The coding system categorizes the social tags into a few categories such as 1.exact match with LCSH 2. Partial match 3. Bibliographical information 4. user specific information. He has found that 16% of the social tags match exactly with LCSH terms while38% of the social tags partially match with LCSH terms, 3% reflected bibliographical information and 43% were user specific information. 46% of the social tags did not match with LCSH terms. He also found that frequency of matching terms is more compared to non matched terms.

Samanta and Rath (2020) studied about the social tags in Library Thing and controlled vocabularies in Sear's List of Subject Heading in the domain of history.

The sample includes 1000 titles and that a small of portion of social tags overlap with SLSH terms and applied the spearman's rank correlation test to find the terminological association of the both. In order to find out the similarity and dissimilarity of highly used social tags and subject headings, Jaccard similarity co-efficient was applied. He concluded that social tags when combined with controlled vocabularies in library catalogues can enhance accessibility of resources. He commented that a social tag contains less subject terms and more personal terms and suffers quality issues.

The main objective of the study conducted by Katagi and Gala (2020) on social tag is to find out the differences and similarities in the tags given on the Library Thing website and subject headings given by OCLC FAST for select books written by Mahatma Gandhi. The authors have analyzed to find the exact synonyms, near synonyms, broad and narrow tags, unique tags, standalone tags. irrelevant tags and the top five most repeated tags. They have concluded that tags are richer in representing the core of the text while controlled vocabulary is universal and offers more precision in finding required resources. Hence no system as such is perfect, but can complement each other.

Gerolimos (2013) have presented a review of literature on the implementation of tagging process in library catalogues. In the beginning the author has examined that most of the research papers and case studies about tagging in libraries include

comparison between tags and subject headings and only a few papers analyze the use of tags in repositories and websites. He has argued that success of social tags in libraries depends on the participation of the users on subject description. The author has concluded that incorporating social tags is essential for libraries but it should never degrade the quality of the subject descriptions.

Methodology

This study has been conducted in the context of the literature reviewed. Many studies have come out comparing social tags and subject heading. This is also a comparative study of social tags in Library thing website (www.librarything.com) and OCLC fast subject headings available at the website of OCLC classify: an experimental classification web service (http://classify.oclc.org). But here we are trying to find out the extend of application of tagging in two different subjects ,i.e. literature(fiction) and chemistry (nonfiction).

For conducting the study, first ten books each from the subject chemistry (DDC No 540) and English literature (DDC No 820) has been collected from the library and catalogued into library thing by creating an account. Before cataloguing, it is ensured that the titles are already available in both the websites. After cataloguing, the social tags and OCLC subject headings are copied into a text

editor with their frequencies and those social tags with frequency number one are avoided from the count.

Subject headings and social tags of each book is then copied to the excel sheet to identify the common terms and differentiate the unique terms. Here if a subject term appears in both library thing and OCLC classify, then it is identified as common term and if a term appears only in library thing it is identified as unique term. Those terms which are not relevant to subject under study and more related to the user's context such 'reading', 'must read', 'owned', 'catalogued' etc are treated as irrelevant term and those social tags such as 'fiction', 'classic', 'modernism' 'English literature','novels' has no relation to the subject headings but they can be considered as relevant terms for retrieval of documents. Hence they are also included under the relevant terms group for analysis.

Analysis of collected data from literature books

Ten books in literature (call no 820) have been selected for the study. The total number of social tags and OCLC subject headings has been brought together in the pattern mentioned in the methodology part. The user assigned tags are ten times more in number than subject headings. Total number of social tags for ten books are 8126 (excluding single frequency tags) where as subject headings are 78 only. (Table 1:- Comparison of social tags to subject heading in literature)

From the table we can interpret that book 3 received highest number of tags (1658) where as book 1 received the lower number, only 64 tags. It is also seen that except book 1 and book 4, all other books have tags above 500. While OCLC fast subject headings given in OCLC classify remains static between 7 and 8 in number.

LITERATURE BOOKS	TOTAL NUMBER OF SUBJECT HEADING S	TOTAL NUMBER OF SOCIAL TAGS	COMMON TERMS	RELEVANT SUBJECT TERMS	USER RELATED TERMS
Book 1	7	64	2 (3%)	40 (63%)	22 (34%)
Book 2	8	893	7 (.8%)	232 (26%)	654 (73%)
Book 3	8	1658	7(.4%)	435 (26%)	1216 (73%)
Book 4	7	146	4 (3%)	79 (54%)	63 (43%)
Book 5	8	629	5(.8%)	307 (49%)	317(50%)
Book 6	8	805	5 (.6%)	340 (42%)	460 (57%)
Book 7	8	1378	8(.6%)	402 (29%)	968 (70%)
Book 8	8	777	7 (.9%)	294 (38%)	476 (61%)
Book 9	8	1004	8 (.8%)	382 (38%)	614 (61%)
Book 10	8	772	8 (1%)	494 (64%)	270 (35%)
TOTAL	78	8126	61 (.75%)	3005 (37%)	5060 (62%)

Table 1:- Comparison of social tags to subject heading in literature

From the comparative table, it is identified that the percentage of common terms is less than one percent in seven books , one percent in one book (book 10) and 3%

in two books (book 1 and 4). The total percentage of common terms is only .75%. The relevant subject terms comes to 37% and user specific terms forms the major percentage i.e. 62%. To be more specific we find that only three books (book 1, 4 and 10) have relevant subject terms higher than user specific terms,(i.e. 43%, 54% and 64% respectively) and in all other books the percentage of relevant subject terms are less than user specific terms.

The analysis shows that user related terms forms the highest percentage (62%) and common terms are forms the lowest percentage (.75%)

Analysis of collected data from Chemistry books

Here also ten books from chemistry (Call no 540) have been taken for study. In chemistry the scene of social tags is different from literature. The number of social tag is less than ten and to the maximum twenty five. The comparison of social tags to subject headings is illustrated in the table 2 .

CHEMISTRY BOOKS	TOTAL NUMBER OF SUBJECT HEADINGS	TOTAL NUMBER OF SOCIAL TAGS	COMMON TERMS	RELEVANT SUBJECT TERMS	USER RELATED TERMS
Book 1	1	2	1 (50%)	1(50%)	0
Book 2	1	4	0	0	4 (100%)
Book 3	2	1	0	0	1 (100%)
Book 4	1	1	1(100%)	0	0

Book 5	**5**	**2**	**1 (50%)**	**1(50%)**	**0**
Book 6	**1**	**1**	**0**	**1(100%)**	**0**
Book 7	**2**	**5**	**1 (20%)**	**2 (40%)**	**2(40%)**
Book 8	**1**	**10**	**1 (10%)**	**1 (10%)**	**8 (80%)**
Book 9	**1**	**23**	**1 (4%)**	**5 (22%)**	**17 (74%)**
Book 10	**4**	**3**	**2 (66%)**	**1 (33%)**	**0**
TOTAL	**19**	**52**	**8 (15%)**	**12 (23%)**	**32 (62%)**

Table 2:- Comparison of social tags to subject heading in Chemistry

The total number of social tags for ten books is 52 (single frequency tags excluded) and total number of subject heading is 19 only. Total percentage of common terms with subject headings is 15% only. Only seven out of the ten books have common terms. Book 2,3 and 6 do not have common terms. 23% of the social tags are subject related terms. Book 2,3,and 4 do not have subject related terms

62% of the social tags are user related terms which is higher than common terms and subject related terms. Only five books have user related terms and still it is the highest percentage. (Book 1, 4,5,6 and 10 do not have user related terms)

Analysis shows that user related terms forms the highest percentage (62%) and common terms forms the least percentage (15%)

Comparison of social tags in chemistry and literature

Table 3 represents a comparison between number of social tags in literature and chemistry. The total number of social tags in literature for 10 books is very high (8126) while in chemistry the total number of social tags for 10 books is 52. User related tags forms the highest percentage in both the cases. The percentage of common terms , i.e. the terms similar to subject headings in OCLC classify is highest in chemistry (15%), while in literature it is only .75%.

SUBJECT	TOTAL NUMBER OF SUBJECT HEADINGS	TOTAL NUMBER OF SOCIALTAGS	COMMON TERMS	RELEVANT SUBJECT TERMS	USER RELATED TERMS
Literature	**78**	**8126**	**61 (.75%)**	**3005 (37%)**	**5060 (62%)**
Chemistry	**19**	**52**	**8 (15%)**	**12 (23%)**	**32 (62%)**

Table 3:- Comparison of social tags in chemistry and literature

Social tag with high frequency of use

In literature the social tag which has highest frequency of use is “fiction” in all the 10 books. While in chemistry the social tag which has highest frequency is “chemistry” in 6 Books i.e. in more than 50% of the books the frequently used social tag is chemistry. In literature the social tag which has highest frequency is not similar to subject headings given in OCLC classify. But in chemistry, the social

tag that has highest frequency of use is similar to the subject headings of OCLC classify.

Application of social tags in fiction and non fiction

From the above analysis we can infer that social tags are more applicable in fiction (literature) than in nonfiction science subjects like chemistry. The reasons are

1. The total number of social tags in literature is 100 times more in chemistry
2. The percentage of common terms to subject headings in literature is only less than one percent (0.75%) while in chemistry it is fifteen percent (15%)
3. The common social tag that has maximum frequency of use in literature is not similar to subject headings while in chemistry it is similar to subject heading and it is the subject term itself.
4. More access points are provided in catalogues for literature books than science books

Findings

1. The social tags which are similar or exact match to subject headings are very poor in literature when compared with the total number of social tags. But in

chemistry the percentage of exact match with subject headings is very high when compared with the total number of social tags.

2. The percentage of user related tags or irrelevant subject tags are high for both the subjects. The users are providing their own access terms for easy identification. This will make the tag cloud bulky and unmanageable. There are also spelling mistakes at certain cases.

3. The total number of social tags in literature is 8126 and in chemistry it is only 52.

4. The subject term which has high frequency of use in literature is “fiction” and in chemistry it is the subject name itself.

5. Social tags are suitable for fiction than nonfiction.

Conclusion

This study is only a demo of the comparison of social tags in two subjects. Here we have identified that social tags are more suitable in fiction. Online library catalogues should make provision to incorporate social tags also into their subject headings. An authority file incorporating the relevant subject tags should be prepared in advance and the unrelated or user related tags should be eliminated from that. The users are incorporating the bibliographical details of the book, native of the author like Indian author, Irish author, other works of the same

author, country of origin etc such terms may be eliminated to maintain quality of the subject descriptions. Libraries should definitely take steps to incorporate social tags into the subject headings in online catalogues without deteriorating the quality and standard of the catalogues.

References

1. Marinho, Leandro Balby, et al. (2012). Recommender systems for social tagging systems. New York: Springer.
2. Melissa De Fino (2011) OCLC Classify, Technical Services Quarterly, 28:4, 456-457, DOI: 10.1080/07317131.2011.598070. Retrieved from https://www.tandfonline.com/action/showCitFormats?doi=10.1080%2F07317131.2011.598070
3. Praveenkumar Vaidya and Harinaryana, N S (2016). The role of social tags in web resource discovery: an evaluation of user-generated keywords. Annals of library and information studies 63(December) p289-297.
4. Praveenkumar Vaidya and Harinarayana, N.S. (2018). Social Semantics and Similarities from User-generated Keywords to Information Retrieval: A Case Study of Social Tags in Marine Science. DESIDOC Journal of Library & Information Technology, 38(1) p.11-15.
5. Rahman, A. I. M. Jakaria. (2012). Social tagging versus Expert created subject headings.. DOI 10.13140/RG.2.1.2426.0968 Retrieved from https://www.researchgate.net/publication/280495267_Social_tagging_versus_Expert_created_subject_headings.
6. Kalyan Sundar Samanta and Durga Sankar Rath (2020) User-Generated Social Tags Versus Librarian-Generated Subject Headings: A Comparative

Study in the Domain of History . DESIDOC Journal of Library & Information Technology, 40(3) p. 176-184.

7. Sachin Katagi and Bhakti Gala (2020) Social Tags of Select Books Written by Mahatma Gandhi : A Comparative Study of Library Thing Tags and OCLC Fast Subject Headings DESIDOC Journal of Library & Information Technology, 40 (1) p. 34-39.
8. Michalis Gerolimos (2013) Tagging for Libraries: A Review of the Effectiveness of Tagging Systems for Library Catalogs, Journal of Library Metadata, 13:1, 36-58, DOI:10.1080/19386389.2013.778730
9. Pennington, Diane Rasmussen and Spiteri, Louise F, Eds. (2019). Social tagging in a linked data environment. London: Facet publishing.
10. Emanuel, Jenny (2009), Next generation catalogs what do they do and why should we care? Reference and users services quarterly, 49(2), p.117-120. Retrieved from https://www.jstor.org.
11. Moulaison, Heather Lea. (2008) Social tagging in the web 2.0 environment: author Vs user tagging. Journal of library metadata 8(2), p101-111. Retrieved from DOI: 10.1080/10911360802087325.